Out[1]:

Out[2]: Click here to toggle on/off the Python code.

PHYS 10792: Introduction to Data Science

2018-2019 Academic Year

Course instructors: Rene Breton (http://www.renebreton.org) - Twitter @BretonRene (https://twitter.com/BretonRene)
Marco Gersabeck (http://www.hep.manchester.ac.uk/u/gersabec) - Twitter @MarcoGersabeck (https://twitter.com/MarcoGersabeck)

Chapters 7+8

Syllabus

1. Probabilities and interpretations
2. Probability distributions
3. Parameter estimation
4. Maximum likelihood + extended maximum likelihood
5. Least square, chi2, correlations
6. Monte Carlo basics
7. **Probability**
8. **Hypothesis testing**
9. Confidence level
10. Goodness of fit tests
11. Limit setting
12. Introduction to multivariate analysis techniques

Topics

7 Probability

- 7.1 Axioms of probability
- 7.2 Empirical probability
- 7.3 Bayesian statistics
- 7.4 Subjective probability
- 7.5 Limitations

8 Hypothesis testing

8.1 Decision making

- 8.1.1 Introductory examples
- 8.1.2 Hypotheses
- 8.1.3 Alternative hypotheses
- 8.1.4 Types of errors
- 8.1.5 Significance and power

7 Probability

7.1 Axioms of probability

Recall from Week 2:

When repeating a measurement the result may change in an unforeseable manner. This is the characteristic of a random system. The degree of randomness can be quantified with the concept of probability.

Let us define the probability following Kolmogorov (1933).

We have a set of possible results $S = E_1, E_2, \ldots$. To each subset A of S, $A \subset S$, one assigns a real number $P(A)$, called probability and satisfying the *axioms of*

We have a set of possible results $S = \{E_1, E_2, ...\}$. To each subset A of S, $A \subset S$, one assigns a real number $P(A)$, called probability and satisfying the axioms of *probability*:

1. For each subset A in S, $P(A) \geq 0$
2. For all disjoint subsets A and B (i.e. $A \cap B = 0$, null intersection), $P(A \cup B) = P(A) + P(B)$ (i.e. the union of the two is simply the sum of the datasets)
3. $P(S) = 1$

The following properties can be derived from these axioms, where the complement to the set of results A, i.e. not A, is denoted by $\overline{A}$:

- $P(\bar{A}) = 1 - P(A)$
- $P(A \cup \overline{A}) = 1$
- $0 \leq P(A) \leq 1$
- $P(\emptyset) = 0$
- If $A \subset B$ then $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Alternative way of the same

A single possible result is often called event. All of the above can then be written for a single event E_i as the subset A. With S being the set of all possible events, the third axiom becomes:

3. $P(S) = \sum P(E_i) = 1$

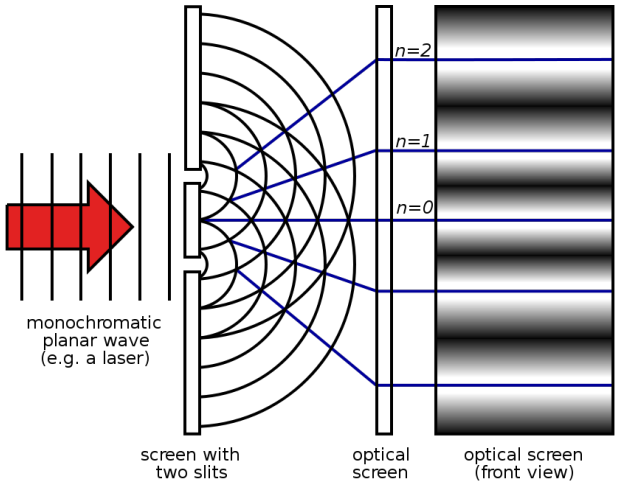
7.2 Empirical probability - The limit of a frequency

Consider an experiment that is executed N times. The outcome A (this could be a single event or a set of events as discussed above) occurs in M of these cases. As $N \to \infty$, the ratio M/N tends to a limit, which is defined as the *probability* $P(A)$ of A.

The experiment may be repeated N times sequentially or N identical experiments may be carried out in parallel. The set of all N outcomes is called *collective* or *ensemble*.

Example: Repeating one experiment

An example for repeating one experiment is the double-slit experiment in which the same double slit is bombarded many times with particles and a distribution builds up on the screen. This distribution corresponds to the probability of observing a particle at a given place on the screen.



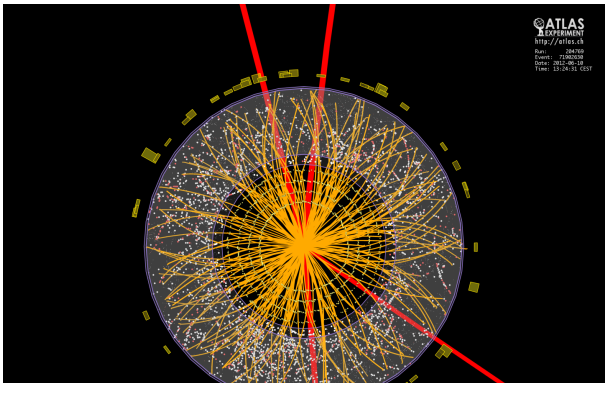
Source Wikimedia/inductiveload (https://commons.wikimedia.org/wiki/File:Two-Slit_Experiment_Light.svg)

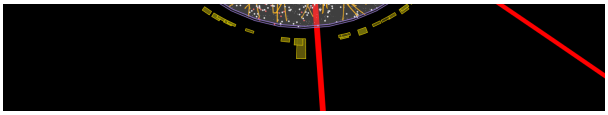
Example: Many independent experiments

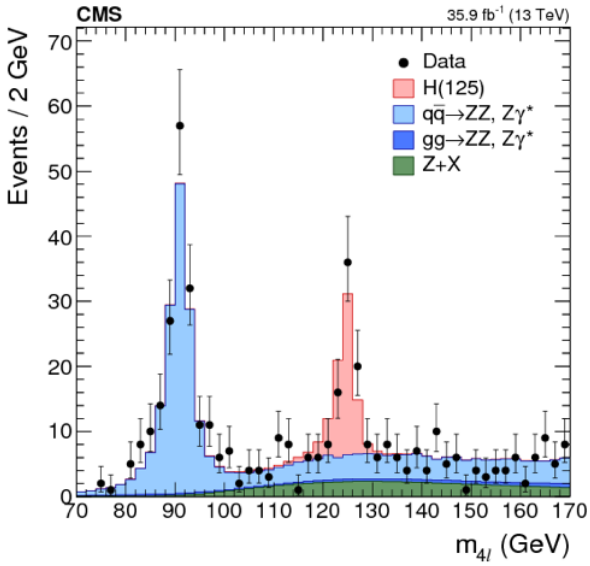
In particle physics colliders produce millions of particle collisions per second.

Each of these can be considered as an independent experiment.

What is studied in the end is the outcome of the ensemble.







Sources CERN/ATLAS (https://cds.cern.ch/record/1459496), CMS, JHEP 1711 (2017) 047 (http://inspirehep.net/record/1608162)

Experiment: Random numbers

Over to menti.com: **31 06 43**

Example: Insurance statistics

The following is a classic example by von Mises.

It is found by the German insurance companies that the fraction of their male clients dying when aged 40 is 1.1%. However, we cannot say that a particular Herr Schmidt has a probability of 1.1% of dying (or 98.9% of surviving) between his 40^{th} and 41^{st} birthdays. The probability of 1.1% refers to all German insured men. Different sample groups that Herr Schmidt may belong to (e.g. all German men, all men, all Germans, all German insured non-smoking men, all German hang-glider pilots) would give different probabilities of his passing away prematurely. Hence, the probability depends on the individual **and** on the collective to which it is considered to belong.

7.3 Bayesian statistiscs

In Bayesian statistics we defined the conditional probability (see Week 2).

The conditional probability $P(A|B)$ is the probability of an event A given that B is true. This is useful in defining the collective in the previous example.

For example the probability that it is Thursday is $P(\text{Tuesday}) = 1/7$, but the probability of it being Tuesday given that you are attending this lecture is $P(\text{Tuesday} \mid \text{DataSci}) = 1/2$ as it only takes place on Tuesdays and Thursdays.

Bayes' theorem states that

$$P(A|B)P(B) = P(A \text{ and } B) = P(B|A)P(A),$$

and hence

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

It is often helpful to express $P(B)$ in terms of whether event A is true or not (with $\overline{A}$ denoting 'not A' as before), which gives

$$P(B) = P(B|A)P(A) + P(B|\overline{A})P(\overline{A}) = P(B|A)P(A) + P(B|\overline{A})[1 - P(A)],$$

and hence by inserting in the previous equation

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})[1 - P(A)]}.$$

Going back to our example, we can write

$$P(\text{DataSci} \mid \text{Tuesday})P(\text{Tuesday}) \qquad 1 \times 1/7 \qquad 1$$

$$P(\text{Tuesday} \mid \text{DataSci}) = \frac{P(\text{DataSci} \mid \text{Tuesday})P(\text{Tuesday})}{P(\text{DataSci} \mid \text{Tuesday})P(\text{Tuesday}) + P(\text{DataSci} \mid \text{not Tuesday})\,[1 - P(\text{Tuesday})]} = \frac{1 \times 1/7}{1 \times 1/7 + 1/6 \times 6/7} = \frac{1}{2}.$$

7.4 Subjective probability

Bayes' theorem can be applied in a way to interpret how a given result strengthens (or weakens) the degree of belief in a given theory:

$$P(\text{theory} \mid \text{result}) = \frac{P(\text{result} \mid \text{theory})}{P(\text{result})} P(\text{theory}).$$

The subjective part lies in the assignment of the probability of the theory being true $P(\text{theory})$.

The interpretation is as follows: if a given result is forbidden by a theory, i.e. $P(\text{result} \mid \text{theory}) = 0$, then its observation disproves the theory, i.e. $P(\text{theory} \mid \text{result}) = 0$. Similarly, if the result is predicted to be unlikely by the theory, its observation reduces the degree of belief in the theory.

If, on the other hand, a result is predicted to be highly likely by the theory, it can strengthen the degree of belief in the theory. However, there are two cases to consider, for which it is useful to consider the previously discussed replacement:

$$P(\text{theory} \mid \text{result}) = \frac{P(\text{result} \mid \text{theory})}{P(\text{result} \mid \text{theory})P(\text{theory}) + P(\text{result} \mid \text{not theory})[1 - P(\text{theory})]} P(\text{theory}).$$

If a result is equally likely regardless of whether or not the theory is true, i.e. $P(\text{result} \mid \text{theory}) = P(\text{result} \mid \text{not theory})$, there is no information gain as this results in $P(\text{theory} \mid \text{result}) = P(\text{theory})$.

The other extreme is that the result is much more likely to occur if the theory is true, i.e. $P(\text{result} \mid \text{theory}) \gg P(\text{result} \mid \text{not theory})$, which leads to the observation of the result being highly predictive as $P(\text{theory} \mid \text{result}) \approx 1$.

Example: Honest Harry

You toss a coin three times and find it showing heads each time. You repeat this as part of a bet with Honest Harry, the used car salesman, and arrive at the same result, which means you lose the bet. The first case will likely not raise any doubts whether or not you a re using a biased, i.e. double-headed, coin. The second case may make you significantly more suspicious. In both cases the probability of the results, given that the coin is unbiased is

$$P(\text{3h} \mid \text{!bias}) = (1/2)^3 = 0.125.$$

We can now calculate the probability of the theory that the coin is biased, given the result of three heads in both cases, $P(\text{bias} \mid \text{3h})$. All that is required is the subjective belief in the theory of the coin being biased. Let us assign $P_{\text{rndm}}(\text{bias}) = 10^{-6}$ for the first case, i.e. that we randomly picked a biased coin, and $P_{\text{Harry}}(\text{bias}) = 0.05$, i.e. that the probability of Honest Harry having made us play with a biased coin is 5%. With the last equation above we now get

$$\begin{aligned} P_{\text{rndm}}(\text{bias} \mid \text{3h}) &= \frac{P(\text{3h} \mid \text{bias})}{P(\text{3h} \mid \text{bias})P_{\text{rndm}}(\text{bias}) + P(\text{3h} \mid \text{!bias})[1 - P_{\text{rndm}}(\text{bias})]} P_{\text{rndm}}(\text{bias}) && (1)\\ &= \frac{1}{1 \times 10^{-6} + 0.125 \times (1 - 10^{-6})} \times 10^{-6} && (2)\\ &= 8 \times 10^{-6}, && (3) \end{aligned}$$

and

$$\begin{aligned} P_{\text{Harry}}(\text{bias} \mid \text{3h}) &= \frac{P(\text{3h} \mid \text{bias})}{P(\text{3h} \mid \text{bias})P_{\text{Harry}}(\text{bias}) + P(\text{3h} \mid \text{!bias})[1 - P_{\text{Harry}}(\text{bias})]} P_{\text{Harry}}(\text{bias}) && (4)\\ &= \frac{1}{1 \times 0.05 + 0.125 \times 0.95} \times 0.05 && (5)\\ &= 0.296. && (6) \end{aligned}$$

The result was able to increase the belief in the theory that Honest Harry had introduced a biased coin because the observation is rather predictive with $P(\text{3h} \mid \text{bias})/P(\text{3h} \mid \text{!bias}) = 8$.

7.5 Limitations

Suppose you measure the mass of the electron as $m = (520 \pm 10)\ \text{keV}/c^2$, i.e. you have obtained a result of $520\ \text{keV}/c^2$ with a resolution of $10\ \text{keV}/c^2$. Assuming a Gaussian resolution function it follows for the measurement of m with resolution σ and the true electron mass m_e

$$P(m|m_e) \propto e^{-(m-m_e)^2/2\sigma^2},$$

which corresponds to $P(\text{result} \mid \text{theory})$.

$P(m)$, corresponding to $P(\text{result})$, is the probability of measuring a given mass, which should be a constant if the measurement apparatus is unbiased.

If we now assume that we know nothing about m_e, we could say $P(\text{theory}) = P(m_e) = \text{const.}$, which leads to the proportionality

$$P(m_e|m) = \frac{P(m|m_e)}{P(m)} P(m_e) \propto P(m|m_e) = e^{-(m-m_e)^2/2\sigma^2}.$$

Note that this is only a proportionality and we cannot quantify this without assuming a concrete value for $P(m)$ and $P(m_e)$.

However, we could equally well have said that we know nothing about the measure of m_e^2, which would alter that interpretation. Both approaches are in principle valid; once more it is vital to be aware of all assumptions and to communicate them in their entirety. This will be dealt with in a more rigorous way in Chapter 9: Confidence levels.

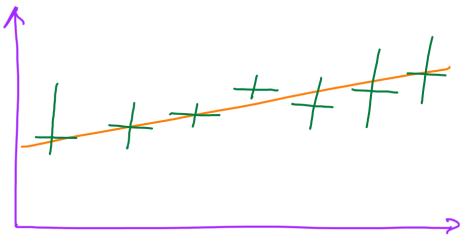
8 Hypothesis testing

8.1 Decision making

Hypothesis testing is essntially decision making. Any test boils down to a yes-or-no answer. This means it is fundamentally different to asking what the measured value of a given observable is as that can in principle take an infinite number of values. Rather, hypothesis taking can evaluate whether the measured value of an observable represents a certain behaviour or not.

Example

Rather than asking what the value of the slope of a fit like the one shown below is, we can ask whether the slope indicates that the linear function is increasing. This is not yet a very precise question and we will pin down the various loose ends as we progress through this lecture.



8.1.1 Introductory examples

In general we need the following to conduct a hypothesis test:

- The assertion that some hypothesis is true,
- A numerical test that is to be applied to data, and
- A hypothesis that is accepted or rejected depending on the outcome of the test.

There are several examples that will all feature in the remainder of this course:

- The interpretation of experiments
- Goodness of fit tests
- Two-sample problems
- Analyses of several samples

Any hypothesis test will not be infallible. We need to choose a level of confidence at which to take the decision. Going back to our example of the linear fit, this level of confidence will be linked to the measured value and its uncertainty. We will discuss how to quantify this in the following.

8.1.2 Hypotheses

Hypotheses are statements that are either true or false. We distinguish simple and composite hypotheses.

Simple hypotheses define the probability distribution function completely. Example hypotheses are:

- These data are drawn from a Poisson distribution of mean 3.4.
- The new treatment has identical effects to the old.

Composite hypotheses combine several probability distribution functions. Example hypotheses are:

- These data are drawn from a Poisson distribution of mean greater than 4.
- The new treatment is an improvement on the old.

Referring back to the components of a hypothesis test, a simple hypothesis corresponds to a numerical test featuring a single equality that is being evaluated, while an inequality would be the numerical test of a composite hypothesis.

8.1.3 Alternative hypotheses

In hypothesis tests we often compare to alternative hypothses. These can take many different forms; examples of alternatives to the first hypothesis (Poisson with mean 3.4) are:

- These data are drawn from a Poisson distribution with another given mean, e.g. 4.5
- These data are drawn from a Poisson distribution with any random mean other than 3.4
- These data are not drawn from a Poisson distribution. This needs to be specified further in order to yield a numerical test.

In general, it is crucial to distinguish between one-tailed directional and two-tailed non-directional tests. A two-tailed test refers to the comparison of a test outcome to a value where we don't care of whether the outcome is less than or greater than the value. In the directional test the sign of the difference between test outcome and comparison value is of importance.

8.1.4 Type I/II errors

The outcome of a hypothesis test is the acceptance or rejection of the hypothesis based on the numerical test. However, it may be that the decision taken does not reflect whether or not the hypothesis is actually true. The two cases where there is a mismatch are called Type I and Type II error according to the following pattern:

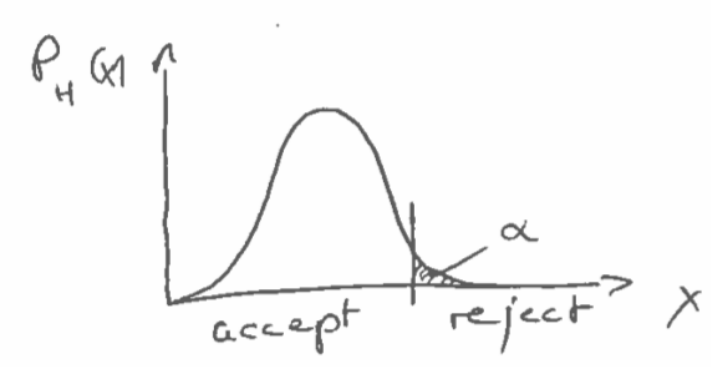
Hypothesis \ Decision	accept	reject
true	:)	Type I error
false	Type II error	:)

8.1.5 Significance and Power

Significance

Type I errors are inevitable and the rate at which they occur is called significance. We know the probabbility ditribution function of the hypothesis, $P_H(x)$, for the case that the test involves the measurement of a quantity x. In our previous example, $P_H(x)$ would be a Poisson distribution with mean 3.4.

The probability distribution function is then divided into a rejection and an acceptance region.



The decision is then taken depending on where the measured value of x falls.

The significance, α, is the integral of the probability distribution of the hypothesis over the rejection region:

$$\alpha = \int_{Reject} P_H(x)dx.$$

Typically, we want α to be small, e.g. 1% or 5%. In reality, we often need to work with the inequality

$$\alpha \geq \int_{Reject} P_H(x)dx,$$

as we may have a range of possible $P_H(x)$ (composite hypothesis) or a discrete distribution for which α cannot be reached exactly. For composite hypotheses we would use the $P_H(x)$ that maximises α. All other incarnations of the composite hypothesis would therefore result in a smaller integral and be equally accepted.

Power

Considering the alternative hypothesis, we can define the integral of the probability distribution of the alternative hypothesis over the acceptance region as

$$\beta = \int_{Accept} P_A(x)dx,$$

or, by integrating of the rejection region as above, we get

$$1 - \beta = \int_{Reject} P_A(x)dx,$$

where $1 - \beta$ is called the power of the test.

Example

Let us consider the following hypothesis:

A data point is drawn from a Poisson distribution with mean less or equal to 5.

We want to test this with 5% significance.

Hence, we need to evaluate

$$0.05 \geq \int_{Reject} Poisson(x; \lambda = 5)dx,$$

as $\lambda = 5$ maximises the integral.

The Poisson distribution is dicrete, so we are looking for the limit, n, of the sum that satisfies

$$0.05 \geq \sum_{x=0}^{n} Poisson(x; \lambda = 5),$$

which is achieved for $n = 9$ with a sum of 0.968.

Therefore, any observed value of 9 or smaller would have the hypothesis accepted.